SANDA CHERATA, TEODOR VUŞCAN, EMMA TĂMÂIANU

SILEX – UN SISTEM LEXICO-MORFOLOGIC COMPUTERIZAT PENTRU ANALIZA TEXTELOR ROMÂNEŞTI

Preocupările în domeniul lingvisticii computaţionale sunt la noi de dată relativ recentă şi nesistematice, astfel că până în prezent ele nu s-au finalizat în instrumente de lucru cum ar fi dicţionarele computerizate, programele de analiză morfosintactică, corectoarele de ortografie etc., instrumente care pentru alte limbi sunt în uz curent de 1–2 decenii. Sistemul lexico-morfologic computerizat (SILEX) pe care îl prezentăm în cele de mai jos – sistem creat de un colectiv de cercetători clujeni – constituie o primă realizare de acest fel în cadrul limbii române. Ea are o valoare apreciabilă în primul rând prin aceea că înlătură cea mai mare parte dintre neajunsurile abordării materialului lingvistic prin metode tradiţionale; spre a măsura utilitatea SILEX-ului, menţionăm câteva dintre aceste neajunsuri: a) timp şi volum de lucru neeconomice, nemaivorbind de faptul că unele cercetări nici nu pot fi realizate prin prelucrări neasistate de calculator; b) incompletitudine a materialului supus cercetării („inventare" deschise, situaţii statistice pe bază de eşantioane nu întotdeauna concludente, clasificări inexacte şi labile); c) descrieri neunitare şi neomogene; d) imprecizia lucrului cu clasicele fişe; e) caracterul „îngheţat" al literei tipărite, fapt care exclude flexibilitatea şi maniabilitatea în exploatare.

Notă terminologică. Lemă = unitate lexicală, reprezentată în mod convenţional printr-o formă de bază (exemplu infinitivul prezent activ al verbului, nominativul singular nearticulat al substantivului etc.) şi apartenenţa ei la o anumită clasă lexico-gramaticală. = procesul prin care fiecare formă ocurentă într-un text este încadrată lemei sale. *Atribut* – termenul este folosit în accepţia sa generală, şi nu în cea specializată din sintaxă.

1. SILEX – prezentare generală

1. 1. *Componente şi funcţii*

SILEX este un produs informatic multifuncţional, astfel conceput încât permite rezolvarea unui spectru larg de probleme din aria cercetărilor de

lingvistică computaţională asupra limbii române şi din cea a prelucrării automate a textelor româneşti.

SILEX este constituit din două componente principale:

a. O componentă statică, de tip *dicţionar computerizat*, care conţine, într-o manieră structurată, un lexic românesc de cca 50 000 de intrări de bază, cu *toate* informaţiile morfologice necesare pentru definirea statutului sistemic şi (gramatical-)textual al unităţilor lexicale (vezi *infra*, **3.**). Dicţionarul computerizat acoperă în întregime informaţia gramaticală cuprinsă in DOOM, verificată, actualizată şi/sau corectată, fără însă a se reduce la aceasta.

b. O componentă dinamică, cu funcţiile de:

(i) generare de forme (pentru cuvintele flexibile); sistemul generează toate formele flexionate ale unui cuvânt, pornind de la lemă şi utilizând atributele asociate ei în dicţionarul morfologic.

(ii) analiză a formelor; analizorul determină clasa lexico-gramaticală a unui cuvânt-ocurenţă, împreună cu valorile categoriilor gramaticale materializate în respectiva formă. În actuala versiune a SILEX, forma este analizată acontextual, adică se realizează toate încadrările posibile (se identifică toate lemele în a căror paradigmă apare forma în cauză)[1]. Din acest motiv, anumite aplicaţii care utilizează SILEX se vor desfăşura interactiv, necesitând, în cazul formelor omografe, selectarea, dintre lemele indicate ca posibile, a celei validate de context.

Exemplu. Pentru forma *a*, analizorul indică lemele:

a avea – verb auxiliar

a – prepoziţie

al – pronume (semiindependent) posesiv.

În articolul de faţă vom prezenta, în datele lui generale, dicţionarul SILEX.

1. 2. *Cerinţe metodologice şi de performanţă*

Având în vedere că un dicţionar computerizat trebuie să conţină, într-o primă etapă, cvasitotalitatea cuvintelor limbii actuale (ceea ce ar însemna aproximativ 100 000 de intrări), se impune o structurare şi reprezentare a informaţiei lingvistice de natură să răspundă cerinţelor de completitudine, coerenţă şi performanţă, în ce priveşte atât spaţiul ocupat, cât şi timpul de acces. Soluţiile pentru structurarea şi reprezentarea informaţiei lingvistice au fost adoptate în funcţie de următoarele condiţii:

(1) orice formă a unui cuvânt trebuie să fie recunoscută, fie direct, având intrare proprie în dicţionar, fie prin mijloace algoritmice eficiente; fiecărei forme trebuie să i se poată ataşa **lema**;

[1] Sunt în curs de elaborare proceduri de restrângere a sferei încadrărilor posibile, proceduri bazate pe analizarea contextului imediat în care apare forma supusă procesului de recunoaştere.

(2) informațiile din dicționar trebuie să permită procesul invers, de generare a întregii paradigme a unui cuvânt, pornind de la lema dată;

(3) timpul de acces la un cuvânt din dicționar să fie cât mai scurt, astfel încât aplicațiile care utilizează dicționarul să se desfășoare fără întârzieri supărătoare;

(4) spațiul de memorie pe care îl ocupă volumul mare al datelor dicționarului să fie cât mai restrâns;

(5) întreținerea dicționarului să se facă prin metode eficiente și simplu de aplicat; aceasta presupune existența facilităților de: introducere a noi cuvinte, corectare, actualizare și îmbogățire a informației cuprinse în dicționar;

(6) structurarea dicționarului trebuie să ofere posibilitatea selectării cuvintelor după toate criteriile lexico-morfologice și după cât mai variate combinații de criterii. Experiența de până acum dovedește că o asemenea facilitate oferă mijloace de mare eficiență atât pentru studii statistice asupra lexicului, cât și pentru verificarea corectitudinii informației din dicționar.

1. 3. *Aplicații ale* ***SILEX***

Tratarea computerizată a textelor românești nu se poate realiza în absența unui instrument cum este SILEX. Funcțiile acestuia, precum și aplicațiile pe care SILEX le face posibile prezintă atât relevanță teoretic-descriptivă, cât și interes practic. Enumerăm doar câteva dintre aceste aplicații, în ordinea crescândă a complexității lor funcționale:

a. dicționar ortografic și morfologic computerizat al limbii române, ușor de întreținut și îmbogățit, furnizabil atât în formă computerizată, cât și în formă tipărită;

b. corector ortografic și morfologic pentru textele românești;

c. sistem pentru studii statistice asupra lexicului limbii române, după cele mai diverse criterii și combinații de criterii;

d. suport pentru orice tip de cercetare sincronică (și, în perspectivă, și diacronică) asupra limbii române (exemple de asemenea obiecte de cercetare: productivitatea anumitor procedee derivative, ponderea relativă a diverselor tipuri de paradigme);

e. sistem pentru studii de statistică lexicală și gramaticală asupra textelor literare;

f. sistem de realizare a concordanțelor pentru operele literare românești, cu lematizare în mare parte automată;

g. suport didactic pentru studierea asistată de calculator a gramaticii limbii române în învățământul preuniversitar și pentru învățarea limbii române ca limbă străină (ortografie, morfologie și lexic).

2. Elaborarea sistemului

SILEX a fost conceput ca proiect interdisciplinar, în cadrul unei colaborări ample între un colectiv de la S.C. Software ITC S.A și Centrul de Analiză a

Textului de la Facultatea de Litere a Universităţii „Babeş-Bolyai".

> SILEX a fost elaborat de cercet. şt. pr. I Teodor Vuşcan şi cercet. şt. pr. II Sanda Cherata (S.C. Software ITC S.A.), Centrul de Analiză a Textului asigurând asistenţa în problemele de descriere lingvistică, prin prof. univ. dr. Marian Papahagi (coordonare) şi asist. univ. Emma Tămâianu.

Realizarea sistemului într-un timp relativ scurt (aproximativ 6 luni) a fost în mare măsură posibilă graţie experienţei în domeniul lingvisticii computaţionale deja acumulate de colectivul de informaticieni, angajat de mai mulţi ani într-un proiect vizând traducerea automată prin intermediul limbii esperanto.

2. 1. *Surse*

În proiectarea dicţionarului computerizat al limbii române s-a plecat de la DEX, DOOM şi GA. Este însă esenţial să precizăm că informaţia morfologică din sursele sus-menţionate nu a putut fi pur şi simplu preluată ca atare, ea nefiind nici unitară, nici completă; în anumite cazuri s-a impus chiar corectarea erorilor de descriere lingvistică şi integrarea unor soluţii propuse şi validate în lucrări de specialitate mai recente[2].

2. 2. *Principii de structurare a informaţiei*

Din punct de vedere abstract, dicţionarul este o mulţime de articole, fiecare articol fiind asociat unei leme. În SILEX, articolele de dicţionar conţin două categorii de informaţii: a) informaţii ce permit determinarea atributelor morfologice ale unei forme flexionate din paradigma lemei respective; b) informaţii care permit regăsirea oricărei forme flexionate din paradigma lemei asociate, precum şi generarea întregii paradigme.

Atributele comune tuturor articolelor sunt:

(1) clasa lexico-gramaticală a lemei; valorile corespund clasificării tradiţionale, din ele derivând atributele proprii şi specifice fiecărei clase;

(2) radicalul / radicalii paradigmei, atribut după ale cărui valori sunt ordonate articolele dicţionarului.

(Pentru inventarul de atribute al fiecărei clase lexico-gramaticale, vezi *infra*, 3. 1..)

2. 2. 1. *Optimizări privind intrările de dicţionar*

Pentru a reduce numărul intrărilor de dicţionar, fără a restrânge mulţimea cuvintelor ce pot fi recunoscute, s-a recurs la soluţia de **a nu introduce** următoarele categorii de cuvinte:

> 1) participiile, inclusiv participiile-adjective; se economisesc astfel aproximativ 5 000 de intrări;
> 2) substantivele provenite din infinitivul lung; se economisesc astfel încă aproximativ 5 000 de intrări;

---

[2] Într-un viitor articol vom prezenta câteva asemenea situaţii.

3) substantivele și adjectivele derivate din radical verbal cu ajutorul sufixului *-tor* (exemplu *muncitor*, *muncitoare*, *semănătoare*); se economisesc astfel aproximativ 7000 de intrări;
4) substantivele omografe cu adjective (exemplu: *calmant*, *diagonală*, *tonic*);
5) substantivele, adjectivele și verbele derivate din radical verbal cu prefixele *ne-* și *re-* (exemplu: *a rescrie*, *neînțeles*, *neînțelegere*, *revăzut*);

În plus, pentru substantivele mobile se introduce o singură intrare, corespunzătoare cuvântului la genul masculin (ex: pentru *elev/elevă* se introduce în dicționar numai cuvântul *elev*).

Cuvintele care nu au intrare proprie în dicționar sunt recunoscute pe baza algoritmilor de flexionare. Această soluție are, pe lângă plusul de economicitate, și avantajul – nu mai puțin important – de a reflecta mai fidel *dinamica* derivărilor lexicale.

2. 2. 2. *Structurarea informațiilor referitoare la flexiune*

Datorită specificului limbii române, prezența în dicționar a informațiilor referitoare la flexiune este indispensabilă pentru orice aplicație de prelucrare a textelor românești. Aceste informații permit atât recunoașterea cuvintelor-ocurență, cât și elaborarea rutinelor de flexionare a oricărei forme de bază. Din cauza complexității procedeelor flexionale, în special din cauza modificărilor produse, în cursul flexiunii, în rădăcina / tema cuvintelor, codificarea din SILEX nu a fost operată după criterii propriu-zis lingvistice, ci după criterii pur formale. În consecință, (sub)clasele flexionale, „rădăcinile" și mulțimile de terminații nu coincid în totalitate cu subcategorizările practicate în descrierea lingvistică. Această codificare ține însă exclusiv de organizarea internă a informațiilor din SILEX, astfel că rezultatul final al analizei / generării formelor, singurul care îl interesează pe utilizator, este întru totul coincident cu realitatea lingvistică.

Astfel, un cuvânt din categoria celor flexionale are, din unghiul analizei automate, următoarea formă:

'radical' + 'terminație',

unde (a) 'radical' înseamnă șirul de caractere invariant în cursul flexiunii (pentru întreaga paradigmă sau doar pentru o parte a acesteia), iar (b) 'terminație' înseamnă șirul de caractere ce se adaugă 'radicalului' pentru a obține o formă flexionată a cuvântului.

În consecință, pentru fiecare intrare dicționarul conține 'radicalul' și o serie de trimiteri codificate la listele de 'terminații' prin a căror atașare rezultă paradigma cuvântului dat.

3. Structurile dicționarului SILEX

3. 1. *Structura atributelor pe clase lexico-gramaticale*

În limbajul algebrei relaționale, dicționarul este o reuniune de relații, fiecare relație corespunzînd uneia dintre clasele lexico-gramaticale tradiționale

(substantiv, adjectiv, verb, adverb etc.). În cele ce urmează descriem aceste relații împreună cu schemele lor, cu semnificația atributelor și cu domeniile de valori.

3. 1. 1. *Relația substantivului*

SUBST **(Sinv, cls, gen, defect, lst_ter, set_parad, lema)**, unde:

Sinv: segmentul de cuvânt comun (invariant al) unei părți a paradigmei;

cls: clasa lexico-gramaticală a cuvântului; valoarea acestui atribut este *sbt*;

gen: genul substantivului; domeniul de valori este {*m*, *f*, *n*, *d*}, unde:
- *m* = masculin;
- *f* = feminin;
- *n* = neutru;
- *d* = indică substantivele mobile;

defect: defectivitatea substantivului; domeniul de valori este {*t*, *s*, *p*}, unde:
- *t* – indică substantivele cu paradigmă completă;
- *s* – indică substantivele cu forme numai pentru singular (defective de plural);
- *p* – indică substantivele cu forme numai pentru plural (defective de singular);

lst_ter: clasa flexională a substantivului, specificată printr-un număr asociat listei de terminații;

set_parad: submulțimea formelor paradigmei în care **Sinv** este partea invariantă;

lema: reprezentată prin forma de N/Ac singular nearticulat.

Exemplu:

Sinv	Categ	Gen	Defect	L_t	Set_p	Lema
șorice	*sbt*	*m*	*t*	*14*	*t*	*șoricel*
femei	*sbt*	*f*	*t*	*16*	*t*	*femeie*
tabel	*sbt*	*n*	*t*	*12*	*t*	*tabel*
șef	*sbt*	*d*	*t*	*1*	*t*	*șef*
făin	*sbt*	*f*	*s*	*1*	*s*	*făină*

3. 1. 2. *Relația adjectivului*

ADJECTIV (Sinv, cls, gen, oms, lst_ter, set_paradm, set_paradf, lema), unde:

Sinv: segmentul de cuvânt comun (invariant al) unei părți a paradigmei;
cls: clasa lexico-gramaticală a cuvântului; valoarea acestui atribut este *adj*;
gen: genul adjectivului; domeniul de valori este {*m*, *f*, *n*}, unde:

- *m* – apare la adjectivele care determină numai substantive de genul masculin;
- *f* – apare la adjectivele care determină numai substantive de genul feminin;
- *n* – apare la adjectivele care determină numai substantive de genul neutru;

Precizare la atributul 'gen'. La adjectiv, atributul 'gen' reflectă exclusiv un fapt de normă actuală standard: datorită semnificației lui lexicale, utilizarea adjectivului în cauză este circumscrisă la aceea de determinant al unui număr finit de substantive dintr-un domeniu semantic compatibil; exemplu *ortic*, *isoscel*, specializate ca determinante pentru *triunghi* (n).

oms: omografia cu un substantiv; domeniul de valori este {*, *m*, *f*, *n*, *d*}, unde:

- * – semnifică absența omografiei;
- *m* – indică omografia cu un substantiv masculin;
- *f* – omografia cu un substantiv feminin;
- *n* – omografia cu un substantiv neutru;
- *d* – omografia cu un substantiv mobil;

Precizare la atributul 'oms'. Atributul privește numai omografia cu un substantiv primar sau rezultat prin substantivarea adjectivului, dar interpretat de vorbitorul contemporan drept cuvânt autonom (exemplu *diagonală*).

lst_ter: clasa flexională a adjectivului, specificată printr-un număr asociat listei de terminații;
set_paradm: submulțimea formelor paradigmei de masculin pentru care **Sinv** este partea invariantă;
set_paradf: submulțimea formelor paradigmei de feminin în care **Sinv** este partea invariantă;
lema: reprezentată prin forma de N/Ac masculin singular (nearticulat).

Exemplu:

Sinv	Categ	Gen	Oms	L_t	S_pm	S_pf	Lema
ştiinţific	*adj*	*mfn*	*	*1*	*t*	*t*	*ştiinţific*
solid	*adj*	*mfn*	*n*	*3*	*t*	*t*	*solid*
ortic	*adj*	*n*	*	*1*	*s*	*p*	*ortic*
român	*adj*	*mfn*	*d*	*1*	*t*	*t*	*român*

3. 1. 3. *Relaţia verbului*

VERB **(Sinv, cls, tip_vrb, s_tor, p_ne, p_re, lst_ter, p_ind, p_conj, p_imp, p_mmcp, p_ps, p_np, lema)**, unde:

Sinv: segmentul de cuvânt comun (invariant al) unei părţi a paradigmei;
cls: clasa lexico-gramaticală; în acest caz are valoarea *vrb*;
tip_vrb: tipul gramatical al verbului; domeniul de valori: {*aux*, *cp*, *pr*}, unde:
aux = verb auxiliar;
cp = verb copulativ;
pr = verb predicativ;
s_tor: posibilitatea formării de substantive şi adjective prin sufixare cu *-tor*, pornind de la radicalul verbal; domeniul de valori este boolean.
p_ne: posibilitatea formării unei alte forme verbale (participiu) prin prefixare cu *ne-* ; domeniul de valori este boolean.
p_re: posibilitatea formării unui alt verb prin prefixare cu *re-* ; domeniul de valori este boolean.
lst_ter: clasa flexională a verbului, specificată printr-un număr asociat listei de terminaţii;
p_ind: mulţimea formelor din paradigma de indicativ prezent în care **Sinv** este parte invariantă;
p_conj: mulţimea formelor din paradigma de conjuctiv prezent în care **Sinv** este parte invariantă;
p_imp: mulţimea formelor din paradigma de indicativ imperfect în care **Sinv** este parte invariantă;
p_mmcp: mulţimea formelor din paradigma de indicativ mai mult ca perfect în care **Sinv** este parte invariantă;
p_ps: mulţimea formelor din paradigma de indicativ perfect simplu în care **Sinv** este parte invariantă;

p_np: mulțimea formelor din paradigma modurilor nepersonale (infinitiv, participiu, gerunziu) și a imperativului în care **Sinv** este parte invariantă;

lema: reprezentată prin forma de infinitiv (fără *a*) prezent activ.

Exemplu:

Sinv	Cat	Tor	Prf	L_t	Ind	Cnj	Imp	Ps	Np	Lema
calcul	*vrb*	*t*	*t*	*1*	*t*	*t*	*t*	*t*	*t*	*calcula*
cit	*vrb*	*t*	*t*	*3*	*t*	*t*	*t*	*t*	*t*	*citi*
mer	*vrb*	*f*	*f*	*7*	*t*	*t*	*t*	*t*	*t*	*merge*
stagn	*vrb*	*f*	*f*	*15*	*t*	*t*	*t*	*t*	*t*	*stagna*

3. 1. 4. *Relația cuvintelor neflexibile*

Neflexibil **(Sinv, cls, atribute, lema),** unde:

Sinv: forma invariantă a cuvântului, care în acest caz coincide cu lema;

cls: clasa lexico-gramaticală a cuvântului; domeniul de valori este {*adv*, *cnj*, *prp*, *int*},unde:

adv = adverb;
cnj = conjuncție;
prp = prepoziție;
int = interjecție.

atribute: diverse, în funcție de clasă.
De exemplu, în cazul adverbului, se semnalează dacă este sau nu „cvasiadverb"; în cazul conjuncției, se semnalează dacă este coordonatoare sau subordonatoare etc.

lema: forma de bază a cuvântului, în cazul acesta invariantă.

3. 2. *Structura listelor de terminații*

Listele de terminații sunt referite din dicționar prin numărul asociat listei. Informațiile din dicționar, împreună cu cele conținute în listele de terminații, permit recunoașterea și flexionarea cuvintelor limbii române, precum și un mare număr de derivări lexicale.

Listele de terminații atașate claselor flexionale au structuri specifice fiecărei clase lexico-gramaticale. Astfel, există liste de terminații pentru substantive, liste pentru subparadigma masculină și, respectiv, pentru subparadigma feminină a adjectivelor și liste pentru fiecare mod/timp al verbului.

În cazul în care formei invariante a cuvântului nu i se ataşează nici o terminaţie (terminaţie vidă), faptul este semnalat în listă prin simbolul @.

3. 2. 1. *Structura listelor de terminaţii pentru substantive şi adjective*

Datorită atributelor comune substantivelor şi adjectivelor, structura listelor de terminaţii pentru aceste clase este aceeaşi, cuprinzând următoarele informaţii:

Cat_flex: numărul listei ataşate respectivei clase flexionale; acesta este referit din dicţionar;
nasn: terminaţia pentru forma de N/Ac singular nearticulat;
gdsn: terminaţia pentru forma de G/D singular nearticulat;
napn: terminaţia pentru forma de N/Ac plural nearticulat;
gdpn: terminaţia pentru forma de G/D plural nearticulat;
nasa: terminaţia pentru forma de N/Ac singular articulat;
gdsa: terminaţia pentru forma de G/D singular articulat;
napa: terminaţia pentru forma de N/Ac plural articulat;
gdpa: terminaţia pentru forma de G/D plural articulat.

Exemplu:

C_f	nasn	gdsn	napn	gdpn	nasa	gdsa	napa	gdpa	Ex.
6	*l*	*l*	*i*	*i*	*lul*	*lui*	*ii*	*ilor*	*şoricel*
16	*e*	@	@	@	*a*	*i*	*le*	*lor*	*femeie*
6	@	@	*i*	*i*	*ul*	*ului*	*ii*	*ilor*	*şef*
3	*d*	*d*	*zi*	*zi*	*dul*	*dului*	*zii*	*zilor*	*solid (adj)*

3. 2. 2. *Structura listelor de terminaţii pentru verbe*

Pentru verbe există şase categorii de liste de terminaţii (toate privind, desigur, diateza activă):

a) pentru indicativ prezent;
b) pentru conjunctiv prezent;
c) pentru imperfectul indicativului;
d) pentru mai mult ca perfectul indicativului;
e) pentru perfectul simplu al indicativului;
f) pentru modurile nepersonale şi imperativ.

Primele cinci categorii de liste au, toate, aceeaşi structură şi prezintă următoarele informaţii:

Cat_flex: numărul listei ataşate respectivei categorii flexionale; acesta este referit din dicţionar;
p1s: terminaţia pentru persoana I singular;
p2s: terminaţia pentru persoana a II-a singular;
p3s: terminaţia pentru persoana a III-a singular;
p1p: terminaţia pentru persoana I plural;
p2p: terminaţia pentru persoana a II-a plural;
p3p: terminaţia pentru persoana a III-a plural.

Exemplu:

C_f	P1s	P2s	P3s	P1p	P2p	P3p	Exemplu
1	*ez*	*ezi*	*ează*	*ăm*	*aţi*	*ează*	*calcula (prez)*
1	*am*	*ai*	*a*	*am*	*aţi*	*au*	*calcula (imprf)*
3	*esc*	*eşti*	*eşte*	*im*	*iţi*	*esc*	*citi (prez)*
7	*g*	*gi*	*ge*	*gem*	*geţi*	*g*	*merge (prez)*

Listele de terminaţii corespunzătoare modurilor nepersonale şi imperativului au următoarea structură:

Cat_flex: numărul listei ataşate respectivei categorii flexionale; acesta este referit din dicţionar;
imper: terminaţia pentru modul imperativ, persoana a II-a singular;
inf: terminaţia pentru infinitiv;
part: terminaţia pentru participiu;
grz: terminaţia pentru gerunziu.

Exemplu:

C_f	Imper	Inf	Part	Grz	Exemplu
1	*ează*	*a*	*at*	*ând*	*calcula*
3	*eşte*	*i*	*it*	*ind*	*citi*
7	*i*	*e*	*s*	*gând*	*merge*
15	*ează*	*a*	*at*	*ând*	*stagna*

4. Concluzii

SILEX a presupus nu doar formalizarea şi codificarea unei descrieri lingvistice preexistente, ci şi, în multe privinţe, găsirea unor soluţii descriptive originale, în prezent încorporate lui. Elaborarea instrumentelor şi procedurilor de analiză lexico-morfologică automată scoate însă la iveală şi probleme teoretice care se cer rezolvate. La acestea ne vom opri în articole următoare.

Dicţionarul SILEX – Prezentare sintetică. Dicţionarul computerizat al SILEX conţine cvasitotalitatea cuvintelor de uz general (acoperind aproximativ 95% dintr-un dicţionar cum este DGLR[3]). Aceasta face ca, în prelucrarea textelor reale, numărul de insuccese (cuvinte nerecunoscute din cauza absenţei lor din dicţionar) să fie foarte mic.

În momentul de faţă, dicţionarul computerizat al SILEX conţine aproximativ 31 000 de intrări şi permite recunoaşterea unui număr de aproximativ 51 000 de leme (multiplicat apoi printr-un număr egal cu totalitatea formelor flexionate ale fiecărei leme).

SILEX este implementat pe un calculator compatibil IBM PC 386.

Bazele de date utilizate ocupă un spaţiu pe disc de 1,68 MB, iar întregul sistem SILEX ocupă un spaţiu pe disc de 2,15 MB.

Modul de organizare a informaţiei şi procedurile de analiză permit deja realizarea unor apreciabile performanţe de timp (într-un minut sunt recunoscute aproximativ 1 000 de cuvinte), performanţe la a căror îmbunătăţire se lucrează în prezent.

Apreciem că, datorită facilităţilor cu care este proiectat, SILEX va constitui un element central în orice viitoare aplicaţie de prelucrare a textelor româneşti.

Universitatea „Babeş-Bolyai"
Facultatea de Litere
Centrul de Analiză a Textului
Cluj-Napoca, str. Horea, 31

[3] Vasile Breban, *Dicţionar general al limbii române*, Bucureşti, 1987.